RedPajama: an Open Dataset for Training Large Language Models

Year: 2024 | *Citations: 151* | *Authors: Maurice Weber, Dan Fu, Quentin Anthony*

Abstract

Large language models are increasingly becoming a cornerstone technology in artificial intelligence, the sciences, and society as a whole, yet the optimal strategies for dataset composition and filtering remain largely elusive. Many of the top-performing models lack transparency in their dataset curation and model development processes, posing an obstacle to the development of fully open language models. In this paper, we identify three core data-related challenges that must be addressed to advance open-source language models. These include (1) transparency in model development, including the data curation process, (2) access to large quantities of high-quality data, and (3) availability of artifacts and metadata for dataset curation and analysis. To address these challenges, we release RedPajama-V1, an open reproduction of the LLaMA training dataset. In addition, we release RedPajama-V2, a massive web-only dataset consisting of raw, unfiltered text data together with quality signals and metadata. Together, the RedPajama datasets comprise over 100 trillion tokens spanning multiple domains and with their quality signals facilitate the filtering of data, aiming to inspire the development of numerous new datasets. To date, these datasets have already been used in the training of strong language models used in production, such as Snowflake Arctic, Salesforce's XGen and AI2's OLMo. To provide insight into the quality of RedPajama, we present a series of analyses and ablation studies with decoder-only language models with up to 1.6B parameters. Our findings demonstrate how quality signals for web data can be effectively leveraged to curate high-quality subsets of the dataset, underscoring the potential of RedPajama to advance the development of transparent and high-performing language models at scale.